



兰州大学

## 本科毕业论文

论文题目 (中文)      基于 Deep Koopman 算子网络的  
非线性系统强化学习研究

论文题目 (英文)      Deep Koopman Network Based  
Reinforcement Learning of Nonlinear System

学生姓名      许国欢  
指导教师      赵东东  
学      院      信息科学与工程学院  
专      业      电子信息科学与工程  
年      级      2020 级

兰州大学教务处

## 诚信责任书

本人郑重声明：本人所呈交的毕业论文（设计），是在导师的指导下独立进行研究所取得的成果。毕业论文（设计）中凡引用他人已经发表或未发表的成果、数据、观点等，均已明确注明出处。除文中已经注明引用的内容外，不包含任何其他个人、集体已经发表或未发表的论文。

本声明的法律责任由本人承担。

论文作者签名： 签名                      日                      期： 签名

## 关于毕业论文（设计）使用授权的声明

本人在导师指导下所完成的论文及相关的职务作品，知识产权归属兰州大学。本人完全了解兰州大学有关保存、使用毕业论文（设计）的规定，同意学校保存或向国家有关部门或机构送交论文的纸质版和电子版，允许论文被查阅和借阅；本人授权兰州大学可以将本毕业论文（设计）的全部或部分内容编入有关数据库进行检索，可以采用任何复制手段保存和汇编本毕业论文（设计）。本人离校后发表、使用毕业论文（设计）或与该毕业论文（设计）直接相关的学术论文或成果时，第一署名单位仍然为兰州大学。

本毕业论文（设计）研究内容：

☒ 可以公开

☐ 不宜公开，已在学位办公室办理保密申请，解密后适用本授权书。

（请在以上选项内选择其中一项打“√”）

论文作者签名： 签名

导师签名： 签名

日                      期： 签名

日                      期： 签名

# 基于 Deep Koopman 算子网络的非线性系统 强化学习研究

## 中文摘要

我的摘要

**关键词：** Koopman 算子理论，深度神经网络，强化学习

# **Deep Koopman Network Based Reinforcement Learning of Nonlinear System**

## **Abstract**

My Abstract

**Keywords:** Koopman Operator Theory, Deep Neural Network, Reinforcement Learning

# 目 录

中文摘要 .....	I
英文摘要 .....	II
第一章 绪 论 .....	1
第二章 背景知识.....	2
2.1 Koopman 算子理论 .....	2
2.1.1 非线性系统描述.....	2
2.1.2 Koopman 算子与系统演化 .....	2
2.1.3 Koopman 算子本征函数 .....	3
2.2 强化学习 .....	4
2.2.1 马尔可夫决策过程与贝尔曼方程.....	4
2.2.2 演员-评委强化学习算法.....	5
第三章 Koopman 算子辅助的最大熵强化学习 .....	7
3.1 动作加权的 Koopman 张量 .....	7
3.2 最大熵强化学习最优控制问题 .....	8
3.2.1 熵正则化 .....	8
3.2.2 Soft Actor-Critic 强化学习算法.....	9
3.3 实验结果.....	11
参考文献 .....	12
附 录 .....	13
致 谢 .....	14

# 第一章 绪 论

这是我的绪论 [1]

## 第二章 背景知识

在本章中，首先讨论一下有关的背景理论与算法。介绍一下 Koopman 算子理论 (Koopman Operator Theory)，并讨论 Koopman 算子对于重塑强化学习 (Reinforcement Learning) 中使用的马尔可夫决策过程 (Markov Decision Process) 的重要作用。同时，对于 Koopman 算子理论与深度神经网络 (Deep Neural Network) 之间的关联。

### 2.1 Koopman 算子理论

系统的强非线性是数据驱动建模和控制领域的核心问题之一，包括基于现代强化学习框架所做的工作。Koopman 算子理论 [cite] 为上述问题提出了一种解决方案。在该理论中，非线性系统动力学可以在提升到无限维度希尔伯特空间时变为线性系统，并可以通过一组新的基底对于该线性系统进行观测。升维工作已被证明对于线性化和简化某些具有挑战性的问题具有显著效果，这与机器学习领域中其他的类似努力相一致。

#### 2.1.1 非线性系统描述

我们应该采用不同的形式对不同类型的系统中系统状态  $x$  进行描述。假设系统为确定性自治系统，我可以采用下面的方式，将一个离散动力系统描述为

$$\dot{x} = F(x) \quad (1)$$

其中， $\dot{x}$  表示下一个时刻的系统状态；或者，采取不同的方式，将一个连续动力系统描述为

$$\frac{d}{dt}x(t) = f(x(t)) \quad (2)$$

而在处理实际问题时，我们通常考虑的是离散动力系统。所以我们可以将连续系统通过流映射算子 (Flow Map Operator) 归纳为一个离散系统，系统演化如下

$$x(t+\tau) = F_\tau(x(t)) = x(t) + \int_t^{t+\tau} f(x(s))ds \quad (3)$$

#### 2.1.2 Koopman 算子与系统演化

Koopman 算子提供了解决非线性系统控制问题的一个新的着眼点。在形式上，我们考虑一个实值向量测量函数  $g: M \rightarrow \mathbb{R}$ ，且都由无限维希尔伯特空间的元素组成，其中  $M$  是一个流形。通常，这个流形被认为是  $L^\infty(X)$ ， $X \subset \mathbb{R}^d$ 。一般情况下，函数  $g$  被称为可观测函数。Koopman 算子理论中指出，Koopman 算子  $\mathcal{K}$  和 Koopman 生成器  $\mathcal{L}$  都是作用于上述

观测函数  $g$  的无限维线性算子，在确定性系统中，有

$$\mathcal{K}g = g \circ F \quad (4a)$$

$$\mathcal{L}g = f \cdot \nabla g \quad (4b)$$

Koopman 生成器  $\mathcal{L}$  和 Koopman 算子有如下的关系

$$\mathcal{L}g = \lim_{t \rightarrow 0} \frac{\mathcal{K}g - g}{t} = \lim_{t \rightarrow 0} \frac{g \circ F - g}{t} \quad (5)$$

Koopman 算子理论可以更广泛地应用于任何马尔可夫过程，但本文以随机性连续时间系统为例，此时，Koopman 算子的定义如下：

$$\mathcal{K}g = \mathbb{E}(g(X)|X_0 = \cdot)$$

$$\mathcal{L}g = \lim_{t \rightarrow 0} \frac{\mathcal{K}g - g}{t}$$

Koopman 算子将测量函数  $g$  沿着路径  $x$  向前演化如下：

$$\mathcal{K}_\tau g(x_t) := g(F_\tau(x_t)) = g(x_{t+\tau}) \quad (6)$$

其中  $F$  代表着系统的演化规律，或者更一般的，在随机自治系统中， $F$  被如下定义为条件预测算子：

$$\mathcal{K}g(x_t) = \mathbb{E}[g(X_{t+\tau})|X_t = x_t] \quad (7)$$

在上述离散系统算子中，普遍使用  $\mathcal{K} := \mathcal{K}_1$ ，在本文中也照此用法。

### 2.1.3 Koopman 算子本征函数

上文提出，可观测函数  $g$  的观测，都存在于无限维的希尔伯特空间中（被称作观测空间），被如 Koopman 算子  $\mathcal{K}$  等无限维的算子推动沿着给定非线性动力学系统演化。因此不难发现，我们可以应用 Koopman 算子理论研究，对于非线性系统的研究，通过观测空间状态线性演化实现。

同时，由于难以捕捉无限维希尔伯特空间中所有可观测函数的演化，所以应该试图识别随着非线性系统动力学而线性演化的关键观测函数，Koopman 算子的本征函数就可以作为一组特殊的观测函数 [cite]：

$$\mathcal{K}\Phi(x_k) = \lambda\Phi(x_k) = \Phi(x_{k+1}) \quad (8)$$

此时，本征函数就会成为观测空间的一组基底，由此，数学上，观测函数应当表示为这组基底的线性组合，如下：

$$g(x) = \sum_{i=1}^n a_i \Phi_i(x) \quad (9)$$

此时，观测空间可被称为状态字典空间，同时，空间的基底可被称为字典函数



当前，从数据中挖掘信息并获得 Koopman 本征函数是现代动力学系统研究的主流方法，被称为数据驱动（Data-Driven）的 Koopman 算子。由此，我们可以通过数据驱动的方式得到 Koopman 算子的本征函数，以得到非线性系统在高维空间中的全局线性表示。

此外，Koopman 算子已经被广泛运用于受控系统。在受控确定性离散时间系统中，我们有：

$$x' = F(x, u) \quad (10)$$

在受控连续时间系统中：

$$\frac{d}{dt} = f(x(t), u(t)) \quad (11)$$

## 2.2 强化学习

强化学习（Reinforcement Learning）是机器学习（Machine Learning）和控制理论（Control Theory）的交叉领域，在强化学习中，智能体学习如何与复杂环境进行交互，并以获得更高的奖励为目标。最近，深度强化学习（Deep Reinforcement Learning）被证明能够在若干项具有挑战性的任务中，实现人类水平或超人类的表现，包括玩电子游戏 [cite] 和策略游戏 [cite]。深度强化学习也越来越多地用于科学和工程应用，包括药物发现 [cite]、机器人操作 [cite]，自动驾驶 [cite] 和无人机赛车 [cite]、流体流量控制 [cite] 和融合控制 [cite]。

### 2.2.1 马尔可夫决策过程与贝尔曼方程

马尔可夫决策过程是具有马尔可夫性质的随机过程。

随机过程是指研究对象是随时间演变的随机现象，在随机过程中，随机现象在某一时刻  $t$  的取值是一个向量随机变量，可以用  $S_t$  表示，所有可能的状态组成状态空间  $\mathcal{S}$ 。我们将已知所有历史状态  $(S_1, \dots, S_t)$  时，某一时刻  $t$  的状态  $S_t$  发生的概率用  $P(S_{t+1}|S_1, \dots, S_t)$  表示。而马尔可夫性质则表示，已知当前时刻状态  $S_t$  时，下一时刻状态  $S_{t+1}$  仅与  $S_t$  有关，用  $P(S_{t+1}|S_t)$  表示。而从前一状态经过随机进入下一状态的过程被称为状态转移。

在下文中，我们假设存在一个无限时域的马尔可夫决策过程。我们假设代理跟随随机策略  $\pi(u|x)$ ，表示在已知状态  $x$  的情况下，采取某个特定动作  $u$  的可能性。由此，在离散时间系统中，状态价值函数定义为：

$$V^\pi(x) = \mathbb{E} \left[ \sum_{k=0}^{\infty} \gamma^k r(x_k, u_k) \mid \pi, x_0 = x \right] \quad (12)$$

其中， $r(x_k, u_k)$  表示智能体得到的奖励； $\gamma \in [0, 1]$  表示折扣率，用于避免在无限时域时无限大的奖励。

在这里我们强调强化学习与控制领域的结合，所以本文中考虑使用线性二次最优控制问题。线性二次最优控制是十分经典的控制领域问题，其中线性代表研究的系统动态可以用一组线性微分方程表示，而其成本为二次泛函。形式上，我们考虑有限时间长度，离散

时间的 LQR，假设离散时间线性系统：

$$x_{k+1} = Ax_k + Bu_k$$

其性能指标为：

$$c(x_k, u_k) = x_k^T Q x_k + u_k^T R u_k \quad (13)$$

将系统线性二次性能指标的相反数作为智能体探索时的奖励：

$$r(x_k, u_k) = -c(x_k, u_k) \quad (14)$$

并改写贝尔曼方程可以得到：

$$V^\pi(x) = \mathbb{E} \left[ \sum_{k=0}^{\infty} -\gamma^k c(x_k, u_k) \mid \pi, x_0 = x \right] \quad (15)$$

### 2.2.2 演员-评委强化学习算法

演员-评论家 (Actor-Critic) 强化学习算法是一种结合价值学习和策略学习的强化学习算法，下面分三步介绍 Actor-Critic 强化学习算法。

首先介绍策略学习，在策略学习中，评价一个策略的方法是评估策略在环境中获得的回报的期望，此时目标函数为：

$$J(\theta) = \mathbb{E}_S [V_\pi(S)] \quad (16)$$

其中， $V_\pi$  代表状态价值函数，可以看到在这个目标函数中，求得了状态价值函数有关状态的期望，所以只涉及到策略  $\pi$  的信息。同时易知，这个目标函数该期望越大，表明该策略越好。

注意，在本文中，对于强化学习的研究只涉及深度强化学习，所以在这里使用深度神经网络的方式对策略网络进行测试。所以，问题转化为如下的一个最大化问题：

$$\max_{\theta} J(\theta) \quad (17)$$

很明显，我们只需要使用梯度上升的方式迭代策略网络。参数更新如下：

$$\theta_{new} = \theta_{old} + \beta \cdot \nabla_{\theta} J(\theta_{now}) \quad (18)$$

其中， $\beta$  代表学习率，是需要调整的超参数，梯度的具体表达式如下：

$$\nabla_{\theta} J(\theta_{now}) \triangleq \frac{\partial J(\theta)}{\partial \theta} \quad (19)$$

以上算法被称为策略梯度

在策略学习中，我们想要得到的是当前策略能得到的价值函数，表示如下：

$$V^\pi(s_t) = \mathbb{E} \left[ \sum_{k=t}^N \gamma^{k-t} \cdot r_k \right] \quad (20)$$

在深度强化学习中，我们通过深度神经网络对价值进行近似，算法上，应该使得上述价值网络逼近真实价值函数。一般使用时间差分（Temporal Differencial）算法来更新价值函数，根据贝尔曼方程：

$$V^\pi(s_t) = \mathbb{E}[r_t + \gamma V^\pi(s_{t+1})] \quad (21)$$

等式左边为价值函数处于时间步  $t$  时对于价值的估计，等式右边是对于左边的一个无偏估计，其中包含的  $r_t$  项使得等式右边比左边更加精确，所以应当使得价值网络逼近等式右边的值。参数更新如下：

$$Q^{new}(s_t, a_t) \leftarrow (1 - \alpha)Q^{now}(s_t, a_t) + \alpha[r_t + \gamma Q^{now}(s_{t+1}, a_{t+1})] \quad (22)$$

最后，通过交替执行策略优化和价值优化，可以使得两者的性能逐渐同步提升。策略网络可以获得更好的状态价值，价值网络可以实现更好的价值估计。最后取用策略网络作为最优控制的策略函数就可以完成 Actor-Critic 算法在最优控制问题上的应用。

### 第三章 Koopman 算子辅助的最大熵强化学习

贝尔曼方程及其连续形式汉密尔顿-雅可比-贝尔曼 (Hamilton-Jacobi-Bellman) 方程在强化学习和控制理论中普遍存在。然而，这些方程对于具有高维状态和非线性的系统来说，很快就变得难以处理。本章将讨论数据驱动的 Koopman 算子与马尔可夫决策过程之间的联系，并介绍借由此开发的两种新的强化学习算法来解决上述的问题。

技术上，我们可以利用 Koopman 算子技术将非线性系统提升到新的坐标中，其中动力学变得近似线性。特别是，Koopman 算子能够通过提升坐标中的线性动力学来捕获给定系统的价值函数的时间演化的期望。

#### 3.1 动作加权的 Koopman 张量

我们现在关注如何通过 Koopman 算子在时间上推进观测基底函数，已知当前状态和动作，也就是说，找到下面的一个映射：

$$(x, u) \mapsto \phi(\dot{x}) \quad (23)$$

我们要求观测函数近似地覆盖每个控制变量  $u$  的价值函数的有限维的 Koopman 不变子空间，因此存在矩阵  $K^u \in \mathbb{R}^{d_x \times d_x}$ ，使得  $K^u \phi(x) = \phi(\dot{x})$ 。

给定轨迹数据和状态字典空间  $\phi$ ，为每个动作  $u \in \mathcal{U}$  构建如上的 Koopman 矩阵  $K^u$ 。为此，采取类似于 SINDYc[cite] 中描述的方法，其中状态和动作的字典空间用于预测下一个状态，即  $\Theta(x, u) \mapsto \dot{x}$  而这里的方法不同的地方有两点。首先，不是试图预测下一个状态  $x$ ，而是试图预测下一个字典函数值  $\phi(x)$ 。其次，为了让状态字典空间覆盖每个  $u$  的 Koopman 不变子空间，状态-动作字典在状态和动作上是可分离的。

这里，将状态-动作字典空间建模为乘法可分离的，就像  $\psi(x) \otimes \phi(x)$  然后进一步假设存在线性映射  $\psi(x) \otimes \phi(x) \mapsto \phi(\dot{x})$ 。下面描述如何为任何的  $u$  构建矩阵  $K^u$ 。

令  $\phi : \mathcal{X} \mapsto \mathbb{R}_x^d$  为状态的特征映射 ( $\phi$  的每一个分量都是一个可观测函数)，并且令  $\psi : \mathcal{U} \mapsto \mathbb{R}_u^d$  为控制动作的特征映射。算法要求对于所有  $u \in \mathcal{U}$  得到 Koopman 算子  $\mathcal{K}^u$  的有限维逼近。令  $\mathcal{T}_K \in \mathbb{R}^{d_x \times d_x \times d_u}$  为如图 [ref] 所示的三维向量。对于任何  $u$ ，如下定义：

$$K^u[i, j] = \sum_{z=1}^d \mathcal{T}_K(i, j, k) \phi(u)[z] \quad (24)$$

形式上， $K^u$  是沿着张量  $\mathcal{T}$  第三维的向量积，同时， $K^u$  作为对于 Koopman 算子  $\mathcal{K}^u$  的有限维近似。因此，我们可以通过学习  $\mathcal{T}_K$ ，来最小化字典函数  $\phi$  的误差：

$$\min_{\mathcal{T}_K} \sum_{i=1}^N \|K^{u_i} \phi(x_i) - \phi(x'_i)\|^2 \quad (25)$$

我们可以重写上述目标,使其成为一个规则的多变量线性回归问题,将  $\mathcal{T}_K$  重新排列为一个大小为  $\mathbb{R}^{d_x \times d_x d_u}$  的二维矩阵。令  $M \in \mathbb{R}^{d_x \times d_x d_u}$ , 其中  $M[i, :] \in \mathbb{R}^{d_x d_u}$  是堆叠二维矩阵  $\mathcal{T}_K[i, :, :]$  所得到的向量,如图 [ref] 所示。令  $\psi(u) \otimes \phi(x) \in \mathbb{R}^{d_x d_u}$  为克罗内克积 (Kronecker Product), 所以有:

$$K^u \phi(x) = M(\psi(u) \otimes \phi(x)) \quad (26)$$

从而, 优化问题变成了常规的线性规划问题:

$$\min_M \sum_{i=1}^N \|M(\phi(u) \otimes \phi(x_i)) - \phi(\dot{x}_i)\|^2 \quad (27)$$

只要我们计算出了  $M$ , 我们就可以通过将  $M$  转换回为三维张量  $\mathcal{T}_K$  的方式为任意系统执行的  $u \in \mathcal{U}$  得到标准的 Koopman 算子, 同时也得到了上述对  $\mathcal{K}^u$  的有限维近似  $K^u \in \mathbb{R}^{d_x \times d_x}$

### 3.2 最大熵强化学习最优控制问题

考虑强化学习在最优控制领域的应用, 在实践中最大的问题有以下两点: 首先, 非常高的样本复杂度: 使得智能体在探索中遇到困难; 其次, 缓慢、不稳定的收敛过程: 探索和利用的权衡是策略优化的重点, 特别是在最优控制中, 不合适的样本利用和探索会使得策略过早的收敛到不良的局部最优值。本文中使用最大熵强化学习算法, 通过在训练目标中合理加入熵正则化项。经过训练, 策略可以最大化预期回报和熵。

#### 3.2.1 熵正则化

粗略的说, 熵是一个表示随机变量随机程度的量。具体的说, 如果一个随机变量的概率分布很集中, 那么这个随机变量的熵就低; 如果一个随机变量的概率分布比较分散, 那么这个随机变量的熵就高。

令  $x$  是具有概率质量密度函数的随机变量  $P$ 。  $H$  的熵是根据  $x$  的分布得出的:

$$H(P) = \mathbb{E}_{x \sim P} [-\log P(x)]$$

在熵正则强化学习中, 代理在每个时间步获得与该时间步的策略熵成比例的奖金。由此, 目标策略变为:

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot | s_t))) \right] \quad (28)$$

其中  $\alpha > 0$  是权衡参数。我们现在可以在如上假定中给出略有不同的价值函数。  $V^\pi$  更改为包含每个时间步的熵奖励:

$$V^\pi = \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t (r(s_t, a_t, s_{t+1}) + \alpha H(\pi(\cdot | s_t))) \mid s_0 = s \right] \quad (29)$$

$Q^\pi$  更改为包括除了第一个时间步以外的每个时间步的熵奖励:

$$Q^\pi(s, a) = \mathbb{E}_{\tau \sim \pi} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) + \alpha \sum_{t=1}^{\infty} \gamma^t H(\pi(\cdot | s_t)) \mid s_0 = s, a_0 = a \right] \quad (30)$$

通过这些定义,  $V^\pi$  和  $Q^\pi$  通过以下方程联系起来:

$$\begin{aligned} V^\pi(s) &= \mathbb{E}_{\dot{s} \sim P, \dot{a} \sim \pi} [r(s, a, \dot{s}) + \gamma(Q^\pi(\dot{s}, \dot{a}) + \alpha H(\pi(\cdot | \dot{s})))] \\ &= \mathbb{E}_{\dot{s} \sim P} [r(s, a, \dot{s}) + \gamma V^\pi(\dot{s})] \end{aligned} \quad (31)$$

标准的强化学习最大化奖励之和的期望, 但是在加入熵正则方法的强化学习中, 将会考虑一个更加一般的最大熵目标 [cite]:

$$J(\pi) = \sum_{t=0}^T \mathbb{E} [r(s_t, a_t) + \alpha \mathcal{H}(\pi(\cdot | s_t))] \quad (32)$$

权重参数  $\alpha$  决定了熵项对于奖励的相对重要性, 从而控制了最优策略迭代中的随机性。

### 3.2.2 Soft Actor-Critic 强化学习算法

下面, 介绍如何通过策略迭代公式设计 SAC 算法, 我们将从推导熵策略迭代 (Soft Policy Iteration) 开始, 这是一种在应用了熵正则的方法下学习最优策略的通用算法, 在最大熵框架中交替进行策略评估和策略改进。

在策略评估步骤中, 我们希望根据等式 (32) 中的最大熵目标计算策略  $\pi$  的值。对于固定策略, 可以从任何函数  $Q: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$  开始, 并且使用下面的修改过的贝尔曼方程迭代计算熵动作价值:

$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p} [V(s_{t+1})] \quad (33)$$

在策略改进步骤中, 策略将逼近于新的熵动作价值函数的指数。如下的这种特定的设定可以保证策略的改进 [cite]:

$$\pi_{new} = \arg \min_{\pi' \in \Pi} D_{KL} \left( \pi'(\cdot | s_t) \parallel \frac{\exp(Q^{\pi_{old}}(s_t, \cdot))}{Z^{\pi_{old}}(s_t)} \right) \quad (34)$$

尽管该算法将可证明地找到最优解, 但我们只能在表格情况 (有限个状态和动作) 下以其精确形式执行它。因此, 我们接下来将对连续域进行估计, 其中我们需要依靠函数逼近器来表示动作价值函数, 这种近似的做法产生了一种新的实用算法, 称为 Soft Actor-Critic 算法。

如上所述, 大型连续域要求我们推导出熵策略迭代的实际逼近。为此, 我们将对动作价值函数和策略使用函数逼近器, 并且交替使用随机梯度下降 (Stochastic Gradient Descent) 来优化这两个网络。我们将考虑参数化状态价值函数  $V_\psi(s_t)$ 、熵动作价值函数  $Q_\theta(s_t, a_t)$  和策略  $\pi_\phi(a_t | s_t)$ 。这些网络的参数是  $\psi, \theta$  和  $\phi$ 。下面介绍这些参数向量的更新规则。

设定如下熵状态函数训练目标，以最小化平方残差：

$$J_V(\psi) = \mathbb{E}_{s_t \sim D} \left[ \frac{1}{2} (V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi} [Q_\theta(s_t, a_t) - \log \pi_\phi(a_t | s_t)])^2 \right] \quad (35)$$

其中， $D$  是先前采样状态和动作的概率分布，或者是经验回放。等式 (35) 的梯度可以用下面的无偏估计表示：

$$\hat{\nabla} J_V(\psi) = \nabla_\phi V_\psi(s_t) (V_\psi(s_t) - Q_\theta(s_t, a_t) + \log \pi_\phi(a_t | s_t)) \quad (36)$$

其中，动作是根据当前策略采样的，而不是从经验回放中抽取。可以如下训练熵动作价值函数参数来最小化熵贝尔曼残差：

$$J_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim D} \left[ \frac{1}{2} (Q_\theta(s_t, a_t) - \hat{Q}(s_t, a_t))^2 \right] \quad (37)$$

其中

$$\hat{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim p} [V_{\bar{\psi}}(s_{t+1})] \quad (38)$$

同样可以用随机梯度下降优化：

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(s_t, a_t) (Q_\theta(s_t, a_t) - r(s_t, a_t) - \gamma V_{\bar{\psi}}(s_{t+1})) \quad (39)$$

更新利用了目标值网络  $V_{\bar{\psi}}$ ，其中  $\bar{\psi}$  在更新参数时加权更新，这种方式已被证明可以稳定训练 [cite]。最后，可以通过直接最小化等式 (34) 中的 KL 散度 (Kullback-Leibler Divergence) 来学习策略参数：

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D} \left[ D_{KL} \left( \pi_\phi(\cdot | s_t) \parallel \frac{\exp(Q_\theta(s_t, \cdot))}{Z_\theta(s_t)} \right) \right] \quad (40)$$

首先使用神经网络变换来参数化策略：

$$\mathbf{a}_t = f_\phi(\epsilon_t; \mathbf{s}_t) \quad (41)$$

其中  $\mathbf{t}$  是输入噪声向量，从一些固定分布中采样，例如球面高斯。我们现在可以将等式 (40) 中的目标重写为：

$$J_\pi(\phi) = \mathbb{E}_{s_t \sim D, \epsilon_t \sim N} [\log \pi_\phi(f_\phi(\epsilon_t; \mathbf{s}_t) | s_t) - Q_\theta(s_t, f_\phi(\epsilon_t; \mathbf{s}_t))] \quad (42)$$

我们可以近似等式 (42) 的梯度：

$$\hat{\nabla}_\phi J_\pi(\phi) = \nabla_\phi \log \pi_\phi(\mathbf{a}_t | \mathbf{s}_t) + (\nabla_{\mathbf{a}_t} \log \pi_\phi(\mathbf{a}_t | \mathbf{s}_t) - \nabla_{\mathbf{a}_t} Q_\theta(s_t, \mathbf{a}_t)) \nabla_\phi f_\phi(\epsilon_t; \mathbf{s}_t) \quad (43)$$

其中， $a_t$  由  $f_\phi(\epsilon_t; \mathbf{s}_t)$  评估。

完整的算法见算法 1。

---

**算法 1** Soft Actor-Critic 强化学习算法
 

---

```

1: 初始化参数向量  $\psi, \bar{\psi}, \theta, \phi$ 
2: for 每一次迭代 do
3:   for 每一个时间步 do
4:      $\mathbf{a}_t \sim \pi_\phi(a_t|s_t)$ 
5:      $\mathbf{s}_{t+1} \sim p(s_{t+1}|s_t, a_t)$ 
6:      $D \leftarrow D \cup (\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})$ 
7:   end for
8:   for 每一次梯度 do
9:      $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$ 
10:     $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_\pi(\phi)$  for  $i \in 1, 2$ 
11:     $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$ 
12:     $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$ 
13:   end for
14: end for

```

---

### 3.3 实验结果



## 参考文献

- [1] Tenne R, Margulis L, Genut M e, et al. Polyhedral and cylindrical structures of tungsten disulphide[J]. Nature, 1992, 360(6403):444–446.

## 附 录

这是我的附录这是我的附录

这是我的附录

## 致 谢

这里是致谢页

(我是谁? 兰朵儿开发者: 余航, 致谢我, 查重时一定会重复的, 哈哈, 开个玩笑, 本科生论文不在查重范围, 而且“毕业论文(设计)检测内容主要为毕业论文(设计)的主体部分”)。

毕业论文（设计）成绩表

导师评语

建议成绩 签名

指导教师（签字） 签名

答辩委员会意见

答辩委员会负责人（签字） \_\_\_\_\_

成绩 \_\_\_\_\_

学院（盖章） \_\_\_\_\_

年 月 日